2Л

УДК 681.3:519.711:004.8

О.А. Логинов
Национальный горный университет, г. Днепропетровск, Украина
walter_topol@i.ua

Актуализация мониторинговых данных на основе классификации

Предложена методика актуализации мониторинговых данных, основанная на нахождении степени доверия к показателям точек наблюдения с применением схем классификации. Разработана соответствующая информационная технология актуализации данных. Проведена апробация на данных гидрологического мониторинга (уровня грунтовых вод в опорных скважинах сети наблюдений) Днепропетровской области, показана целесообразность ее применения.

Постановка проблемы в общем виде и ее связь с важными научными и практическими задачами

В связи с возрастающими объемами статистической информации, накапливаемой в распределенных, разрозненных источниках данных, и постоянно меняющимися требованиями к анализу информации актуальным направлением исследований становится актуализация данных для решения аналитических задач.

Актуализация данных – приведение данных в соответствии с состоянием отображаемых объектов предметной области. Актуализация реализуется посредством операций добавления, исключения и редактирования записей [1]. Актуализация данных позволяет определить наборы данных с низкой степенью доверия, выделять заблаговременно «ложные» наборы данных, при прогнозировании снижает вероятность ошибочного результата.

Основным требованием аналитика является достоверность используемой информации. Как правило, анализ предметной области показывает изменение структур данных мониторинга, потребность проведения согласования данных, без чего нельзя говорить о достоверности. Помимо этого данные содержат ошибки измерений. При исследовании мониторинговых данных требуется рассматривать не только отдельные показатели, но и динамические ряды, встает проблема обеспечения минимума методологических искажений и неконтролируемых человеком потерь информации в процессе ее обработки.

Общепринятой методики актуализации данных не существует, в разных предметных областях используются свои методы. В то же время для мониторинговых данных можно предложить универсальный подход, основанный на классификации. Классификация – отношение соответствия между классом и диапазоном изменения показателя [2], [3]. Существует много схем одномерной классификации, в данной работе предлагается применять такие, как «равные интервалы», «естественная разбивка», «стандартное отклонение» [4].

Цель работы – разработка методики и соответствующей информационной технологии актуализации данных с применением схем классификации, проверка их адекватности на реальных мониторинговых данных Днепропетровской области.

Методика актуализации данных с использованием схем классификации

Концептуальная схема разработанной методики актуализации данных и соответствующей информационной технологии приведена на рис. 1.

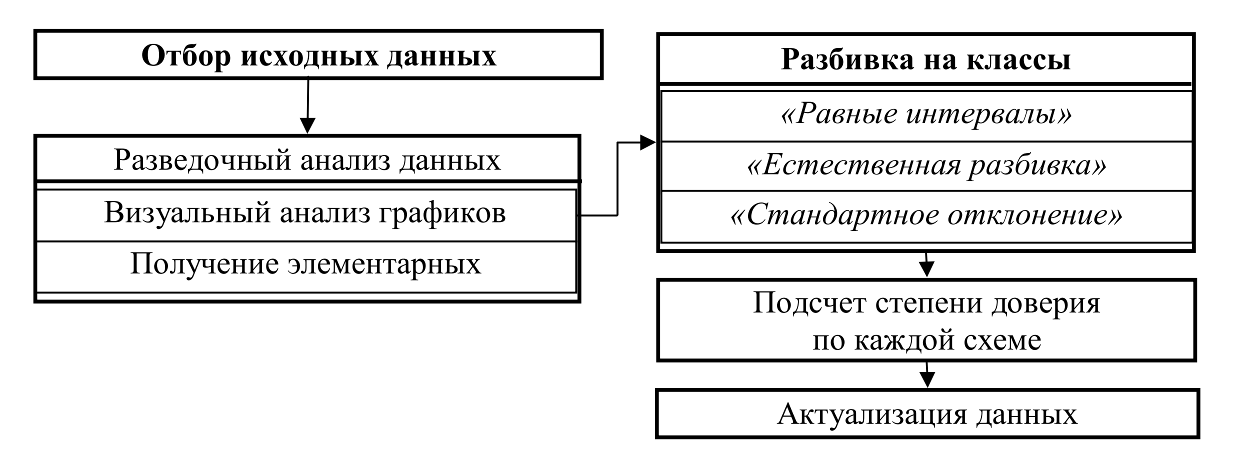


Рисунок 1 – Концептуальная схема методики актуализации данных

Рассмотрим основные этапы получения актуализированных данных (на примере гидрологического мониторинга).

Этап 1. Отбор исходных данных.

Грунтовые воды – подземные воды первого от поверхности земли постоянного водоносного горизонта, не имеющего сверху сплошной кровли водонепроницаемых пород, не обладают напором и подвержены сезонным колебаниям уровня и дебита [5]. Моделирование и прогнозирование уровня грунтовых вод (УГВ) является важной народнохозяйственной задачей.

Полный набор исходных данных составляет ежемесячный мониторинг по 22 скважинам за 31 год наблюдения с 1974 по 2005, всего 8184 значений. Из него отбираются показатели всех скважин за один месяц (март) по всем годам. Анализу подвергается выборка $\{Z_i\}$, $i = 1, 2,\ldots, N$, по $N = 22$ скважинам за $M = 31$ год наблюдений, всего 982 значения (пример исходных данных – табл. 1). $Z_i = (Z_{i,1},\ldots,Z_{i,M})$ – наблюдения по i-й скважине.

Таблица 1 – Выборка по 22 скважинам на 31 год наблюдений

N	N_skv	x	y	1974	1975	...	j	...	2004	2005
1	14360	34,4861	34,4861	3,02	3,13	...	...	...	2,77	1,60
2	14354	34,4906	34,4906	4,22	4,28	...	...	...	4,14	3,07
	...	...	...	...	...	...	...	...	...	...
i	...	...	...	...	...	...	$Z_{i,j}$	...	...	...
	...	...	...	...	...	...	...	...	...	...
21	6337	35,1675	35,1675	1,42	1,54	...	...	...	1,05	1,05
22	8760	34,8933	34,8933	2,12	2,00	...	...	...	2,18	2,18

Этап 2. Разведочный анализ данных.

Получение элементарных статистик (пример – табл. 2), построение графиков, визуальный анализ графиков (рис. 2) по скважинам.

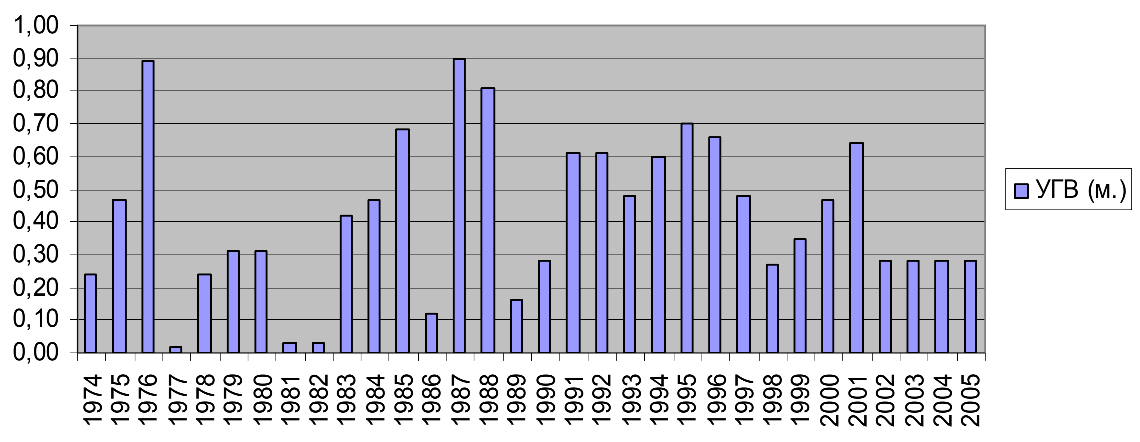


Рисунок 2 – Уровень грунтовых вод, март 1974 – 2005 гг.

Таблица 2 – Элементарные статистики ряда Z_{11}

Максимум	0,90
Минимум	0,02
Среднее	0,42
Дисперсия	0,06
Стандартное отклонение	0,24

Этап 3. Разбивка на классы.

Производится разбивка выборки $\{Z_i\}$, $i = 1,\ldots,N$ на классы по трем схемам классификации: «равные интервалы», «естественная разбивка», «стандартное отклонение».

1. «*Равные интервалы*»:

а) вычисляются $Z_{\min}, Z_{\max}$ – соответственно минимальное и максимальное значения выборки;

б) промежуток $[Z_{\min}, Z_{\max}]$ делится на k интервалов равной длины;

в) элементу ряда присваивается номер класса, соответствующий номеру интервала, в который попадает его значение.

2. «*Естественная разбивка*» (аналог кластеризации):

а) ряд представляется в виде гистограммы;

б) разбивка на k классов проводится в соответствии с резкими скачками в значениях частот; в один класс попадают элементы ряда, имеющие близкие значения;

в) каждому элементу ряда присваивается номер класса.

3. «*Стандартное отклонение*»:

а) вычисляется $\bar{z}$ – среднее значение выборки;

б) вычисляется s – стандартное отклонение;

в) разбиение на интервалы начинается от $\bar{z}$ последовательным прибавлением и вычитанием m-й доли ($0 < m \le 1$) стандартного отклонения s:

$$[\bar{Z} - m \cdot s; \bar{Z} + m \cdot s); \quad [\bar{Z} - 2m \times s; \bar{Z} - m \times s), [\bar{Z} + m \times s; \bar{Z} + 2m \times s); \ \ldots$$

г) каждому элементу выборки, значение которого попало в интервал $[\bar{Z}-m\cdot s; \bar{Z}+m\cdot s]$, присваивается номер класса.

Подсчитывается число попаданий в каждый из классов значений отдельно по каждому, по каждой скважине. Полученные градации для классификации приведены в табл. 3.

Таблица 3 – Градации классификаций

Градации	Естественная разбивка		Равные интервалы		Стандартное отклонение	
I интервал	0,00 - 1,46		0 - 2,12		0,00 - 0,81	
II интервал	1,46 - 2,89		2,12 - 4,25		0,81 - 1,97	
III интервал	2,89 - 4,50		4,25 - 6,37		1,97 - 3,13	
IV интервал	4,50 - 7,39		6,37 - 8,50		3,13 - 4,29	
V интервал	7,39 - 10,62		8,50 - 10,62		4,29 - 5,45	
VI интервал	—		—		5,45 - 10,10	

Пример полученной классификации «естественная разбивка» приведен в табл. 4.

Таблица 4 – Классификация «естественная разбивка»

N	N_skv	x	y	1974	1975	. . .	1986	1987	. . .	2004	2005
1	14360	34.4861	48.0364	3,02	3,13	. . .	1,77	2,52	. . .	2,77	1,60
2	14354	34.4906	48.0417	4,22	4,28	. . .	3,28	4,03	. . .	4,14	3,07
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
21	6337	35.1675	48.4728	1,42	1,54	. . .	0,77	2,29	. . .	1,05	1,05
22	8760	34.8933	48.5228	2,12	2,00	. . .	1,92	2,26	. . .	2,18	2,18

Этап 4. Подсчет степени доверия по каждой схеме.

а) определяются частости $V_{ij},\ i=1,...,N,\ j=1,...,K$ попаданий объектов наблюдений в заданные К классов;

б) находится число попаданий для самого многочисленного класса в Z_i относительно всех попаданий значений скважины по каждой из трех схем разбивки

$$V_i^* = \arg\max_j V_{ij}; \qquad (1)$$

в) находится степень доверия к значениям Z_i

$$\mu(Z_i) = \frac{V_i^*}{M}; \qquad (2)$$

г) находится средняя степень доверия по трем схемам классификации.

Этап 5. Актуализация данных.

Выделяются и признаются не пригодными к дальнейшей работе те показатели скважин, чьи степени доверия ниже заданного порога δ.

Как видно из табл. 5, выделены три скважины с наименьшей (при заданном пороге δ = 0,7) степенью доверия по трем способам разбивки и три скважины по двум способам разбивки («равные интервалы» и «стандартное отклонение»), всего 6 скважин.

Схема «естественная разбивка» выделяет только 3 скважины, они также выделены двумя другими способами разбивки, ее можно считать проверочной.

Таблица 5 – Результат актуализации данных

N	N_skv	x	y	Равные интервалы	Естественная разбивка	Стандартное отклонение	Среднее
1	14360	34.4861	48.0364	0,56	0,75	0,63	0,65
2	14354	34.4906	48.0417	0,84	0,78	0,69	0,77
3	14361	34.2859	48.0150	0,47	0,41	0,41	0,43
4	15064	36.0939	48.2426	0,66	0,59	0,38	0,54
5	14358	34.1467	48.1769	0,97	0,97	0,91	0,95
6	12752	36.0244	48.1772	1,00	1,00	0,91	0,97
7	14699	36.0333	48.1833	1,00	0,72	0,84	0,85
8	14698	36.0281	48.185	0,94	0,75	0,94	0,88
9	12926	36.1522	48.0556	0,88	0,97	1,00	0,95
10	15584	36.5644	48.1106	0,97	1,00	1,00	0,99
11	15000	36.0238	48.2632	1,00	1,00	0,91	0,97
12	5946	36.05	48.31	1,00	1,00	0,56	0,85
13	14329	35.9033	48.5492	0,66	0,91	0,53	0,70
14	15221	35.8394	48.5689	0,50	0,84	0,69	0,68
15	6951	35.5057	48.3031	0,94	0,94	0,84	0,91
16	6966	35.92	48.4619	0,94	0,94	0,72	0,86
17	8776	35.0744	48.9269	0,66	0,53	0,66	0,61
18	8777	35.0803	48.9206	0,88	0,88	0,66	0,80
19	14697	35.0997	48.9219	0,94	0,84	0,94	0,91
20	8519	34.5638	49.1131	1,00	0,88	0,69	0,85
21	6337	35.1675	48.4728	0,97	0,91	0,91	0,93
22	8760	34.8933	48.5228	0,69	0,97	0,72	0,79
				0,47	0,41	0,38	min
				1,00	1,00	1,00	max

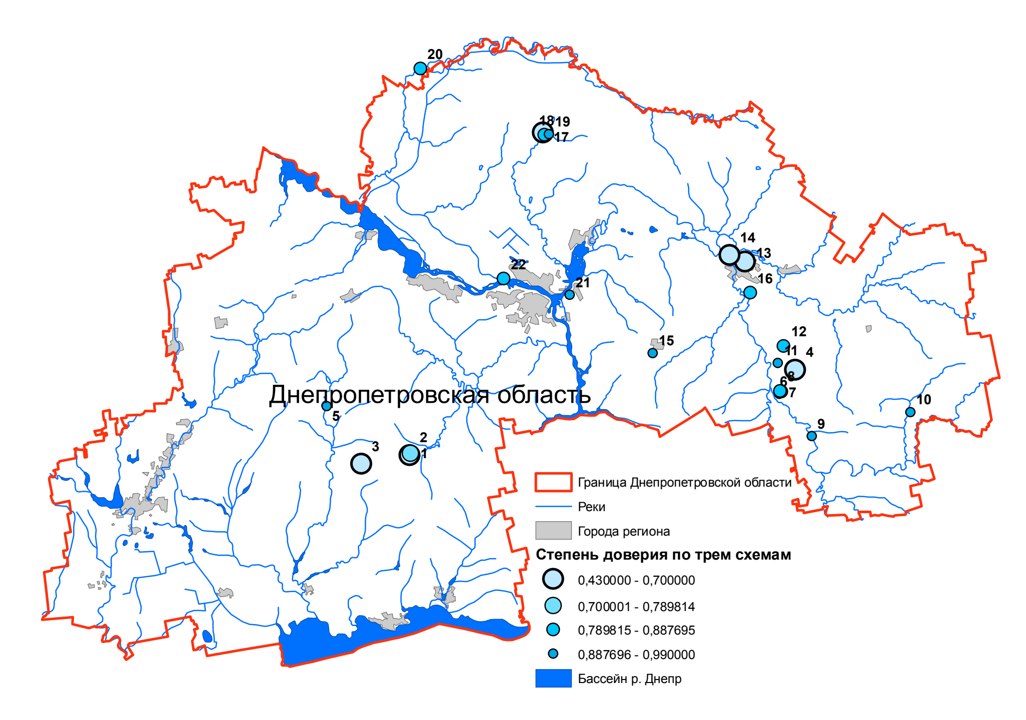


Рисунок 3 – Опорные скважины Днепропетровской области и степени доверия к показателям мониторинга

На рис. 3 приведена карта расположения скважин в Днепропетровской области и средние по трем схемам степени доверия к ним.

Скважины с низкой степенью доверия при прогнозировании могут привести к неадекватным результатам.

Актуализированные данные дают возможность при дальнейшем прогнозировании снизить процент ошибок.

Выводы

Разработана и реализована компьютерная информационная технология актуализации данных, основанная на схемах классификации данных. Дальнейшие исследования пространственно распределенных данных мониторинга целесообразно направить на двумерный вариант классификации и связать результаты с данными геостатистического анализа. При более детальном исследовании гидрологических данных необходимо принять во внимание то, что на территории Днепропетровской области расположено три различных гидрологических бассейна.

Литература

1. Электронный словарь URL [Электронный ресурс]. – Режим доступа : http://www.finam.ru/dictionary/wordf0092500014/default.asp?n=1
2. Сарычева Л.В. Компьютерный эколого-социально-экономический мониторинг регионов. Геоинформационное обеспечение : [монография] / Л.В. Сарычева. – Днепропетровск : НГУ, 2003. – 174 с.
3. Сарычева Л.В. Компьютерный эколого-социально-экономический мониторинг регионов. Математическое обеспечение : [монография] / Л.В. Сарычева. – Днепропетровск : НГУ, 2003. – 222 с.
4. Саричева Л.В. Схеми класифікації регіонів за показниками еколого-соціально-економічного моніторингу в геоінформаційній системі / Л.В. Саричева, О.В. Качанов // Геоінформатика. – 2002. – № 4. – С. 53-63.
5. Рубан С.А. Гідрогеологічна оцінка і прогнози режиму підземних вод України / С.А. Рубан, М.А. Шинкаревський. – Дніпропетровськ : Укрпівденгеологія, 2005. – 372 с.

О.О. Логінов

Актуалізація моніторингових даних на основі класифікації

Запропонована методика актуалізації моніторингових даних, основана на знаходженні ступеня довіри до показників точок спостереження із застосуванням схем класифікації. Розроблена відповідна інформаційна технологія актуалізації даних. Проведена апробація на даних гідрологічного моніторингу (рівня ґрунтових вод в опорних свердловинах мережі спостереження) Дніпропетровської області, показана доцільність її застосування.

O.A. Loginov

Actualization of Monitoring Data on the Basis of Classification

The methodology of monitoring data actualization, based on finding of degree of belief to indexes of observation points with classification scheme application was suggested. Appropriate information technology of data actualization was developed. The approbation on hydrologic monitoring data (level of subterranean waters in reference borehole of observation net) of Dnepropetrovsk region was carried out, the expediency of its application was shown.

Статья поступила в редакцию 10.07.2009.